DOCUMENT RESUME

ED 424 652 EA 029 418

AUTHOR Livingston, Samuel A.
TITLE Results of the Pilot Test of the School Leaders' Licensure Assessment.
PUB DATE 1998-04-00
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Certification; Construct Validity; Educational Administration; Elementary Secondary Education; Instructional Leadership; *Principals; *School Administration; Tables (Data); *Test Validity
IDENTIFIERS *School Leaders Licensure Assessment

ABSTRACT

The School Leaders' Licensure Assessment (SLLA) is a 6-hour, written, constructed-response test intended for use in the licensing of principals, headmasters, and other school leaders. The pilot test was administered in 1996 to 247 students who were within six credits of completing their graduate degree programs in educational leadership. About 36 percent were from urban districts, 32 percent from suburban districts, and 32 percent from rural districts. About 66 percent of the test takers were female, 51 percent were white, 33 percent were black, and 14 percent were Hispanic, 2 percent were American Indian. The scorers of the pilot test were practicing school principals. The scoring was holistic, each scorer read the response to the question and assigned a numerical rating, and each question had a set of explicit scoring rules. The total scores of the participants ranged from 20 to 74, with a mean of 53 and a standard deviation of 11. The largest differences in the subgroups were found between the white participants and those from the minority populations. Tables and figures provide various visual representations of the findings. (RJM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

March 20, 1998

Results of the Pilot Test of the School Leaders' Licensure Assessment[1]

Samuel A. Livingston
Educational Testing Service

ED 424 652

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
S. Livingston
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

EA 029 418

1

March 20, 1998

Results of the Pilot Test of the School Leaders' Licensure Assessment[1]

Samuel A. Livingston
Educational Testing Service

The School Leaders' Licensure Assessment (SLLA) is a six-hour written constructed-response test intended for use in the licensing of school principals, headmasters, etc. It consists of four sections, each containing a different type of question. The total score is a weighted sum of the scores on the four sections; the weights are chosen so that each section accounts for a specified proportion to the maximum possible total score. The range of possible total scores is 0 to 100. Table 1 describes the format of the test.

ETS pilot tested the first form of the SLLA in December of 1996, in Kentucky, Mississippi, Missouri, North Carolina, and Texas. The participants were 247 students in graduate degree programs in educational leadership and administration who were within six credits of completing their degree requirements. Figure 1 presents demographic data describing the group of participants. About one-third were from urban districts, one-third from suburban districts, and one-third from rural districts. About two-thirds were female. About half were White, one-third were Black, and one-seventh were Hispanic. (A small number were American Indian.)

Section III of the SLLA consists of two cases. ETS pilot tested three cases, by creating three versions of the test form being pilot tested. One version included cases 1 and 2; one version included cases 1 and 3; and one version included cases 2 and 3. These three versions were packaged in alternating sequence ("spiraled"), so that the groups of participants responding to the three cases would be randomly equivalent.

The scorers of the pilot test were practicing school principals. The scoring was holistic: the scorer read the response to the question and assigned a numerical rating. Each question had a set of explicit scoring rules, but the application of these rules to the actual responses required judgment on the part of the scorer. For the two cases, the possible ratings were 0, 1, 2, or 3; for all other questions, the possible ratings were 0, 1, or 2. In the pilot testing, two scorers independently scored each response to each question. These two independent responses provided the data for computing statistics describing the inter-rater consistency of the scoring. For all other analyses, the two ratings were averaged, with adjudication of discrepancies of more than one point.[2]

---

[1] To be presented at the annual meeting of the American Educational Research Association, San Diego, CA, on April 17, 1998.

[2] If the two ratings of a response differed by more than one rating point, a third scorer scored the response. If the third scorer's rating was midway between the two original ratings, the two original ratings were allowed to remain. Otherwise, the third scorer's rating replaced the one most different from it.

EA 029 418

Figure 2 shows the inter-rater agreement for the three types of questions rated on a 0/1/2 scale. The proportion of responses receiving the same rating from both raters varied from 78 percent (for the short vignettes) to 68 percent (for the document-based questions). Differences of two points were rare. Figures 3, 4, and 5 show the inter-rater agreement separately for the individual questions in each of these three sections of the SLLA. The proportion of exact agreement varied from .65 to .97 for the short vignettes, from .63 to .79 for the long vignettes, and from .58 to .78 for the document-based questions. These results provided the basis for identifying questions requiring revision, either in the question itself or in the scoring rules.

Figure 6 shows the inter-rater agreement for the three cases, which were rated on a 0/1/2/3 scale. The proportion of responses receiving the same rating from both raters varied from 42 to .86. The proportion of responses for which the ratings differed by no more than one point varied from .90 to 1.00.

To estimate the inter-rater reliability of the total scores, ETS used the first and second ratings to estimate the reliability of scores on each question and combined the estimates with the composite-reliability formula. The estimated inter-rater reliability of the total scores based on two ratings of each response was .93. ETS also estimated the inter-rater reliability of the total scores based on two ratings of the cases but only one rating of the responses to the other questions. The estimated inter-rater reliability of the total scores under this procedure was .90.

ETS' estimate of the alternate-forms reliability of the scores is based on the assumption that, *within each section*, the items included in a particular form of the SLLA are effectively a random sample of all possible items that could have been included.[3] The estimated alternate-forms reliability of the total scores was .79.

Table 2 shows the intercorrelations and the estimated reliabilities of the section scores. The three correlations involving the case section ranged from .48 to .51; the intercorrelations of the other sections ranged from .42 to .45. These correlations are not much lower than the highest correlations consistent with the estimated reliablity of the sectionscores.

The total scores of the 247 participants ranged from 20 to 74, with a mean of 53 and a standard deviation of 11. Figures 7, 8, and 9 are box-and-whisker plots showing the 90th, 75th, 50th, 25th, and 10th percentiles of the distribution of total scores, for all participants and for subgroups defined by sex, location, and race/ethnicity. The largest differences between subgroups were between the White participants and those from the two minority populations, as shown in Figure 9.

---

[3] This assumption implies the use of coefficient alpha to estimate the alternate-forms reliability of the section scores. However, because there were only three cases, taken by different (but overlapping) samples of participants, ETS used a more conservative estimate of the reliability of the case section. The estimation procedure was to compute the intercorrelations of scores on the three cases, choose the smallest of the three correlations, and apply the Spearman-Brown formula (with k = 2, the number of cases in a single operational form of the SLLA).

Were the differences -- or the similarities -- in the performance of the different groups consistent across the four different types of questions? Figures 10, 11, and 12 compare the performance of the subgroups on each section of the SLLA. In general, the between-group differences tended to be fairly consistent across the four types of questions. There were two possible exceptions to this generalization, both involving performance on the cases. In comparison to the other groups, the suburban participants tended to score somewhat higher and the Hispanic participants tended to score somewhat lower on the cases than might be expected from their performance on the other types of questions. However, the data from the cases are averaged over only three different questions. It is not clear that these differences would generalize to other cases or even to other participants from the same demographic subgroups.

Table 1. Format of the School Leaders' Licensure Assessment (SLLA)

Section	Item Type	Number of Items	Time Allowed	Rating scale	Contribution to Total
I	Short vignette	10	1 hour	0 to 2	20 %
II	Long Vignette	6	1 hour	0 to 2	20 %
III	Case	2	2 hours	0 to 3	30 %
IV	Document-Based	7	2 hours	0 to 2	30 %

Table 2. Correlations and reliabilities of section scores

	Short Vignettes	Long vignettes	Document-based questions	Cases
Short vignettes		.42	.45	.50
Long vignettes	.42		.45	.48
Document-based questions	.45	.45		.51
Cases	.50	.48	.51	
Reliability of section scores	.48*	.42*	.59*	.46 - .58**

* Coefficient alpha

** Estimated from correlation of pairs of cases

Figure 1.
Demographic characteristics of the participants.

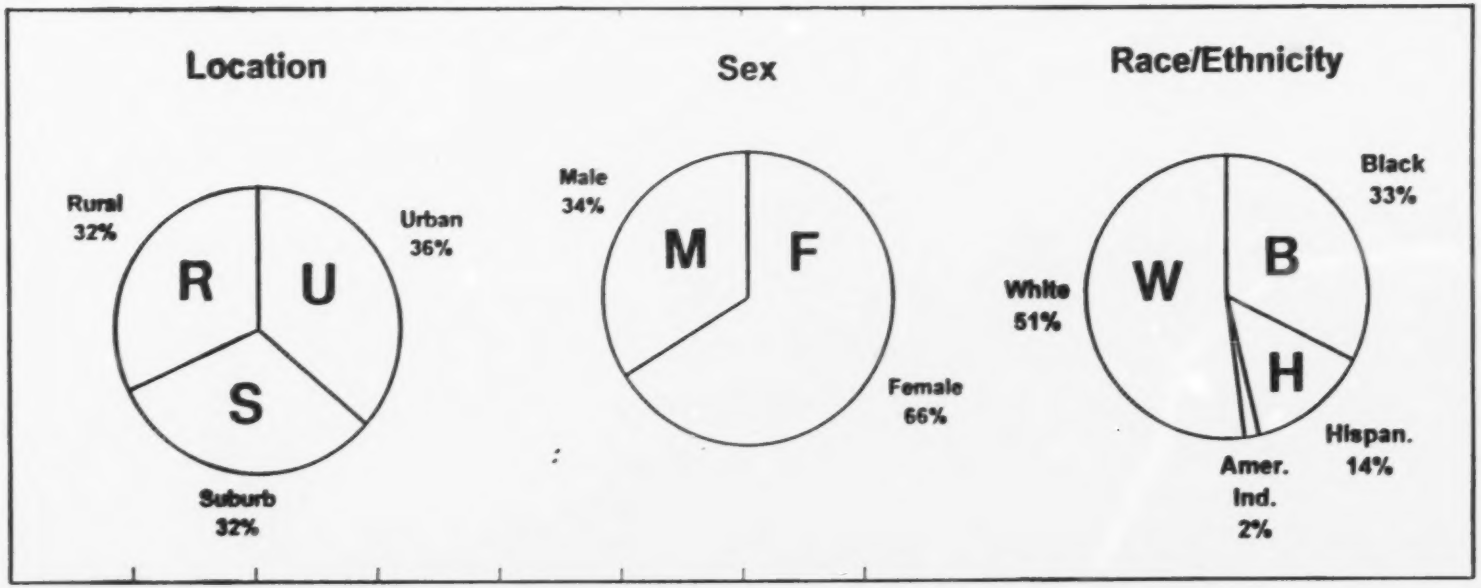


Figure 2. Inter-rater agreement

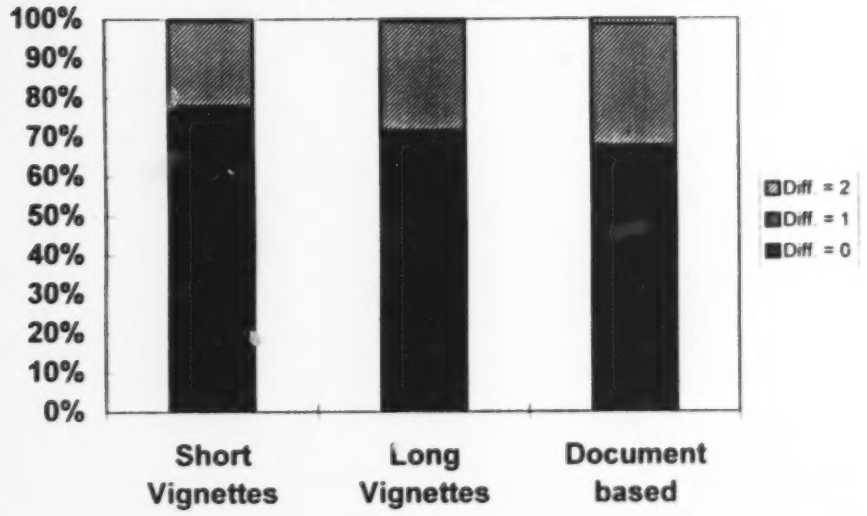


Figure 3. Inter-rater agreement: short vignettes

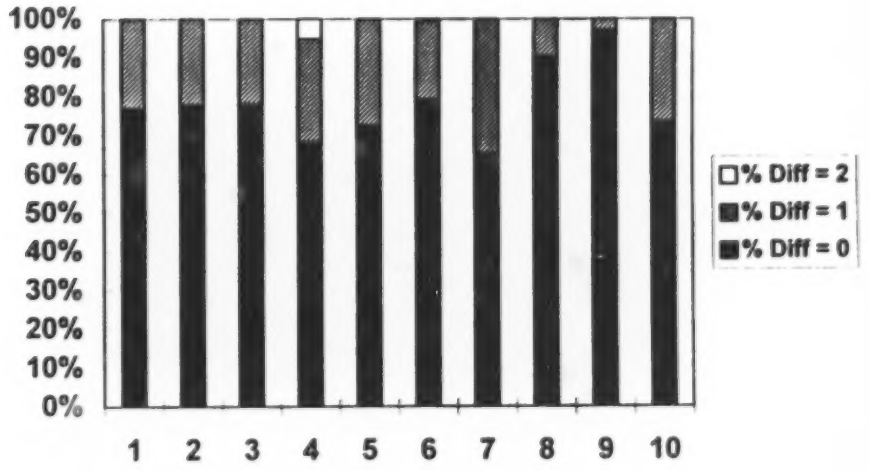


Figure 4. Inter-rater agreement: long vignettes

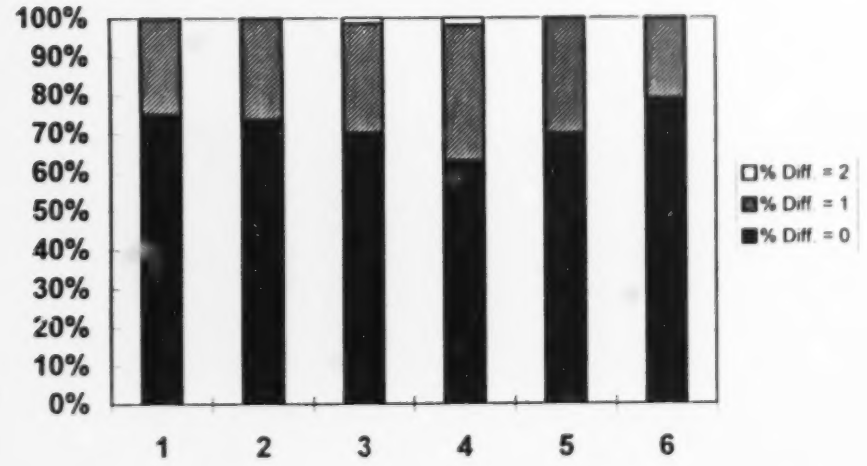


Figure 5. Inter-rater agreement: document-based questions

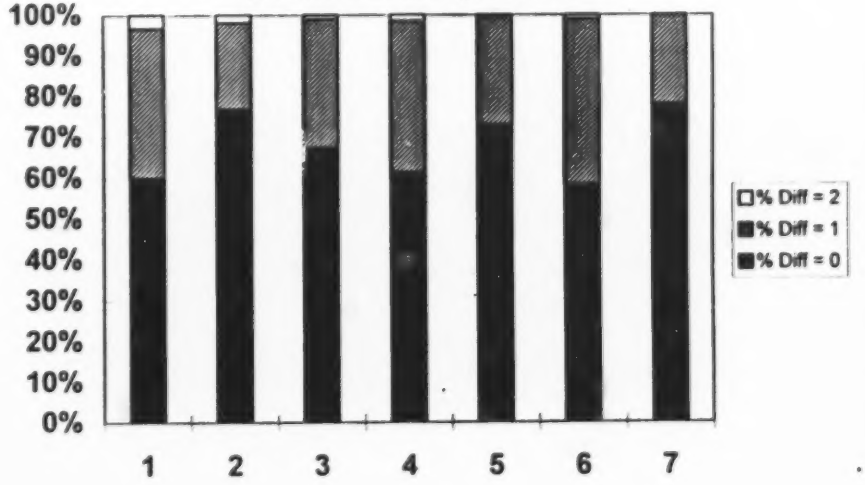


Figure 6. Inter-rater agreement: cases

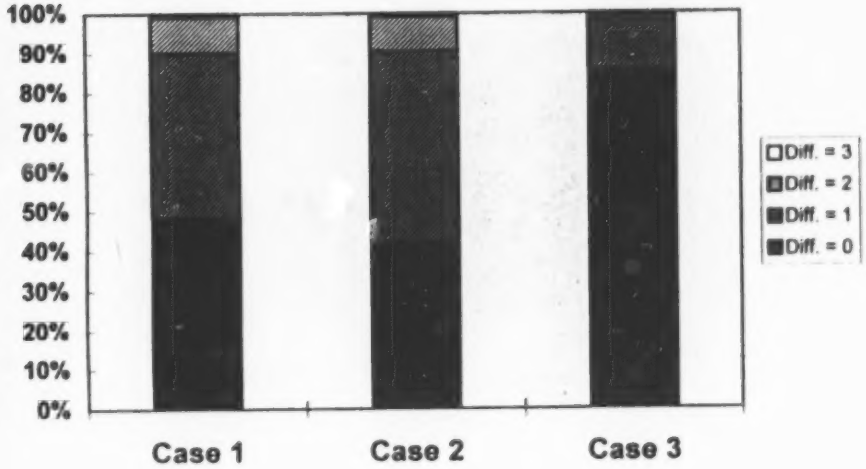


Figure 7. Total score percentiles (90th, 75th, 50th, 25th, 10th): Total Group, Male, Female

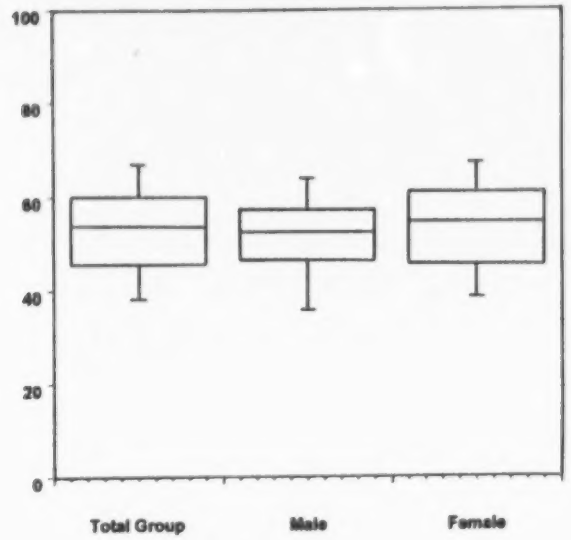


Figure 8. Total score percentiles (90th, 75th, 50th, 25th, 10th): Urban, Suburban, Rural

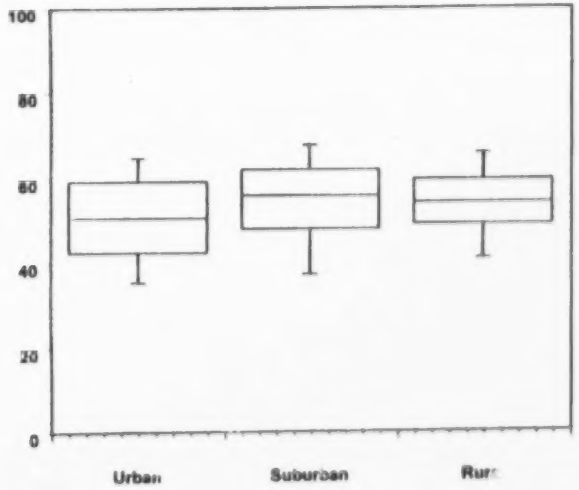
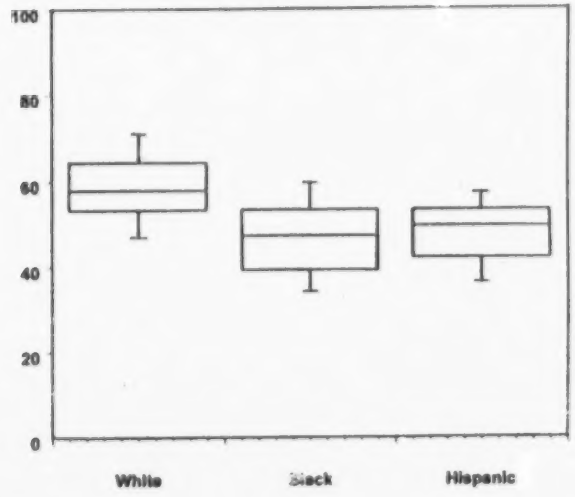


Figure 9. Total score percentiles (90th, 75th, 50th, 25th, 10th): White, Black, Hispanic



BEST COPY AVAILABLE

Figure 10. Mean score as percent of maximum possible: Male, Female

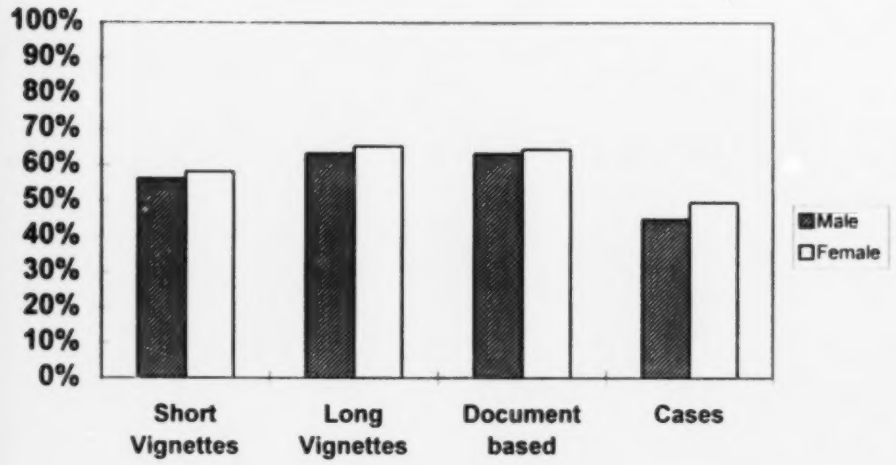


Figure 11. Mean score as percent of maximum possible: White, Black, Hispanic

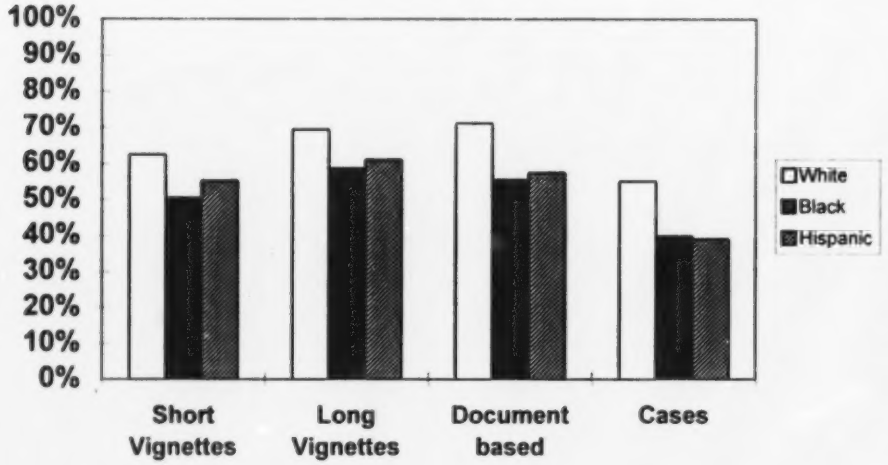


Figure 12. Mean score as percent of maximum possible: Urban, Suburban, Rural